(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENTAMT

(12) **Patentschrift**

(10) **DE 195 01 599 C 1**

(51) Int. Cl.[6]: G 10 L 5/06

(21) Aktenzeichen: 195 01 599.1-53
(22) Anmeldetag: 20. 1. 95
(43) Offenlegungstag: —
(45) Veröffentlichungstag der Patenterteilung: 2. 5. 96

DE 195 01 599 C 1

(73) Patentinhaber:
Daimler-Benz Aktiengesellschaft, 70567 Stuttgart, DE

(74) Vertreter:
Amersbach, W., Dipl.-Ing., 89250 Senden

(72) Erfinder:
Class, Fritz, Dr., 72587 Römerstein, DE; Kaltenmeier, Alfred, Dr., 89075 Ulm, DE; Regel-Brietzmann, Peter, Dr., 89073 Ulm, DE; Kilian, Ute, 72587 Römerstein, DE

(56) Für die Beurteilung der Patentfähigkeit in Betracht gezogene Druckschriften:

US 42 77 644

NEYH, H.: »Automatische Spracherkennung: Architektur und Suchstrategie aus statistischer Sicht«, In: DE-Z: Informatik Forschung und Entwicklung, Bd. 7, H. 2, 1992, S. 83-94;

LOWERRE, B., REDDY, R.: »The HARPY Speach Under- standing System«, In: Readings in peech recog- nition, 1990, Morgan Kaufmann Publishers Inc., ISBN 1-55860-124-4, S. 576-586;

(54) Verfahren zur Spracherkennung

(57) Für die Spracherkennung von in Sätzen verbundenen Wortfolgen wird ein Verfahren vorgeschlagen, welches bei Anwendungen mit definierter Syntax diese in ein Sprachmodell, vorzugsweise ein Bigram-Sprachmodell eines HMM-Erkennungssystems integriert und insbesondere mehrfach auftretende Wörter durch Indizierung eindeutig kennzeichnet. Dadurch ist eine Durchführung des Erkennungsprozesses wie ohne Integration der Syntax-Information möglich.

DE 195 01 599 C 1

Beschreibung

Die Erfindung betrifft ein Verfahren zur Spracherkennung nach dem Oberbegriff des Patentanspruchs 1.

Bei der Erkennung verbundener d. h. fließender Sprache, die eine beliebige Kombination aller Wörter zuläßt, steigt die Fehlerrate im Vergleich zur Einzelworterkennung erheblich an. Um dem entgegenzuwirken, kann beispielsweise Wissen über zulässige Wortfolgen in sogenannten Sprachmodellen gespeichert und bei der Erkennung verwertet werden. Die Anzahl zulässiger Sätze kann dadurch erheblich eingeschränkt werden.

Sprachmodelle werden gebräuchlich als N-gram Modelle definiert, wobei N als Tiefe des Modells bezeichnet wird und die Anzahl aufeinanderfolgender Wörter innerhalb einer Wortfolge angibt, die bei der aktuellen Bewertung einer Wortfolgenhypothese berücksichtigt werden. Wegen des mit zunehmenden N schnell steigenden Aufwands im Erkennungsprozeß wird bevorzugt das mit N=2 besonders einfache Bigram-Modell angewandt, welches nur Zweierkombinationen von Wörtern berücksichtigt. Die Sprachmodelle können weiter vereinfacht werden durch die Zusammenfassung von Wörtern, die im gleichen Kontext auftreten, ohne aber unbedingt dieselbe Bedeutung haben zu müssen, zu Wortgruppen (z. B. alle Wochentage). Statt einzelner Wortübergänge kann in den Sprachmodellen der Übergang von einer Wortgruppe zur anderen betrachtet werden.

In Informatik Forsch. Entw. (1992) 7, S. 8—97, sind Grundprobleme der automatischen Erkennung fließender Sprache eingehend behandelt und Lösungsansätze aus der Sicht der statistischen Entscheidungstheorie beschrieben. Im Vordergrund steht die stochastische Modellierung von Wissensquellen für Akustik und Linguistik z. B. in Form von Phonem-Modellen, Aussprache-Lexikon und Sprachmodell.

Aus "The HARPY Speach Understanding System" in Readings in Speech recognition, 1990, Morgan Kaufmann Publishers Inc. ist ein Spracherkennungssystem mit stark eingeschränkter Anzahl zulässiger Sätze bekannt. Die die Zulässigkeit bestimmenden syntaktischen und semantischen Einschränkungen können in Grammatik-Gleichungen formuliert und als Graph dargestellt werden. Um von der vollständigen, aber mit großem Verarbeitungsaufwand verbundenen Grammatik-Definition zu einem kompakten Sprachmodell mit vertretbarem Verarbeitungsaufwand zu gelangen, werden einige Vereinfachungen eingeführt.

Solche Vereinfachungen sind aber teilweise nur möglich, wenn für das Sprachmodell in Kauf genommen wird, daß in der ursprünglichen Grammatik-Definition unzulässige Wortfolgen wieder als zulässig erscheinen. Bei dem HARPY-System werden schließlich die Wörter durch ihre phonetischen Definitionen ersetzt und so ein phonetisches Modell für einen Ganzsatzerkenner geschaffen.

In der US 4 277 644 ist ein Verfahren zur Spracherkennung beschrieben, das eine begrenzte Menge zulässiger Sätze erkennt. Die die Anzahl zulässiger Sätze begrenzende Syntax ist in Form eines Endlichen Automaten (Finite State Automata, FSA) in einem Wortfolgespeicher abgelegt. Das Einbinden eines Endlichen Automaten in einen Spracherkenner bedeutet einen in Struktur Funktionalität genau auf diesen Automaten ausgerichteten Aufbau des Erkennungssystems.

Aufgabe der vorliegenden Erfindung ist es, ein Verfahren zur Spracherkennung anzugeben, daß bei geringem Verarbeitungsaufwand eine hohe Erkennungsleistung aufweist.

Die Erfindung ist im Patentanspruch 1 beschrieben. Die Unteransprüche enthalten vorteilhafte Ausgestaltungen und Weiterbildungen der Erfindung.

Die Erfindung ermöglicht durch die unterschiedbare Kennzeichnung mehrfach in der Grammatik der Menge der zulässigen Sätze auftretender Wörter im Sprachmodell die zulässigen Vorläufer eines bestimmten Wortes an bestimmter Satzposition implizit vollständig zu erfassen, ohne daß explizit alle zulässigen vorangegangenen Übergänge zu diesem Wort gespeichert werden müssen. Dies entspricht einem N-gram-Sprachmodell mit von der jeweiligen Wortposition abhängigem variablem N. Die unterscheidbare Kennzeichnung mehrfach auftretender gleicher Wörter sei im folgenden als Indizieren der Wörter bezeichnet.

Vorzugsweise kann die Syntaxinformation in einem Bigram-Sprachmodell integriert werden. Der Erkennungsprozeß, der vorzugsweise ein HMM (Hidden Markov Model)-Erkennungsprozeß ist, kann in gleicher Weise ablaufen wie ohne die Integration der Syntax in das Sprachmodell.

Eine wesentliche Erweiterung eines für die akustische Worterkennung herangezogenen gebräuchlichen Aussprachelexikons ist nicht notwendig, da allen im Sprachmodell unterschiedlich indizierten Exemplaren des gleichen Wortes ein und derselbe Lexikoneintrag zugeordnet werden kann. Die Bigram-Syntaxinformation kann dann vorteilhafterweise dadurch berücksichtigt werden, daß dem aus einer Folge von Wortuntereinheiten bestehenden Lexikoneintrag entsprechend dem mehrfachen Auftreten im Sprachmodell mehrere Wortendeknoten zugewiesen werden.

Bei der Spracherkennung nach dem erfindungsgemäßen Verfahren werden eingegebene Sprachsignale immer syntaktisch richtigen Sätzen zugewiesen. Vorzugsweise ist daher die Möglichkeit vorgesehen, daß das Erkennungssystem eine Eingabe zurückweist. Vorteilhaft hierfür ist die Zuweisung eines Wahrscheinlichkeitswerts an erkannte Sätze und Vergleich der Wahrscheinlichkeitswerte mit einer vorgebbaren Rückweisungsschwelle. Die globale Satzwahrscheinlichkeit, normiert auf die Satzlänge, bildet ein gut geeignetes Maß für die Zuweisung der Wahrscheinlichkeitswerte. In die globale Satzwahrscheinlichkeit werden insbesondere die Wahrscheinlichkeiten bei der akustischen Erkennung der einzelnen Wörter einbezogen. Berücksichtigt werden können darüberhinaus auch Wahrscheinlichkeiten aus statistischen Verteilungen von Wortfolgen im Sprachmodell oder Häufigkeiten von Sätzen in Trainingsmengen.

Die Wahrscheinlichkeitsbewertung wird vorzugsweise auch während des laufenden Erkennungsprozesses durchgeführt und als Grundlage für ein Ausblenden von Pfaden mit zu geringer Wahrscheinlichkeit herangezogen.

Die Erfindung ist nachfolgend unter Bezugnahme auf die Abbildungen noch eingehend veranschaulicht.

Die **Fig.** 1a zeigt ein einfaches Beispiel eines Netzwerk-Graphen für ein Sprachmodell, welches aus den Wörtern w1 bis w6 zwischen dem Satzanfang Start und dem Satzende Ende die Wortfolgen w1w3w6, w1w4w6, w2w3w1, w2w5w1 als Sätze zuläßt. Die aus dem Graphen ableitbare Bigram-Information über die zulässigen Nachfolger zu jedem Wort w1 bis w6 ist als Tabelle in **Fig.** 1b angegeben. In einem auf diese Bigram-Information gestützten Sprachmodell erscheinen aber nicht zulässige Sätze wie z. B. w1w3w1w4w6 als zulässig.

Die demgegenüber wesentliche Änderung gemäß der Erfindung ist aus **Fig.** 2a und **Fig.** 2b ersichtlich. Die durch den Netzwerk-Graphen nach **Fig.** 1a festgelegte Menge der zulässigen Sätze enthält die Wörter w1 und w3 jeweils in zwei syntaktisch verschiedenen Positionen. Diese mehrfach vorkommenden Wörter sind nunmehr in **Fig.** 2a als voneinander unterscheidbare Exemplare durch Indizierung gekennzeichnet, wobei der Index _m mit in als ganzzahliger Laufzahl innerhalb des Graphen in an sich beliebiger Reihenfolge auf die Mehrfach-Exemplare eines Wortes vergeben werden kann. Wichtig ist, daß durch die Indizierung Wörter in syntaktischen Positionen, die nicht ohne Änderung der Zulässigkeit aller Sätze vertauscht werden können, eindeutig gekennzeichnet werden. Zur Vereinheitlichung der Notation sind auch alle einmalig auftretenden Wörter mit einem Index_1 versehen. Die Bigram-Informationstabelle in **Fig.** 2b zu dem Graphen von **Fig.** 2a zeigt sich gegenüber der Tabelle in **Fig.** 1b um die Mehrfach-Exemplare erweitert, gibt aber nunmehr eine dem Graphen exakt gleiche Vorschrift über alle zulässigen Sätze wieder und weist einen geringeren mittleren Verzweigungsgrad auf.

Da die phonetischen Repräsentanten für alle Mehrfach-Exemplare desselben Wortes identisch sind, braucht das diese phonetischen Repräsentanten enthaltende Aussprache-Lexikon nicht im gleichen Maße erweitert werden. Es kann für alle Mehrfach-Exemplare desselben Wortes auf denselben Lexikon-Eintrag zurückgegriffen werden, wobei lediglich am Wortende wieder eine eindeutige Zuordnung zu den jeweils zulässigen Nachfolgern ermöglicht werden muß. Hierfür können vorteilhafterweise zu einem betroffenen Lexikon-Eintrag mehrere Wortendeknoten vorgesehen sein, welche die unterschiedlichen Syntax-Einschränkungen der durch Indizierung unterscheidbaren Wortpositionen berücksichtigen.

Bei der vorteilhaften Zusammenfassung von Wörtern zu Wortgruppen treten an die Stelle der Wörter w1_1 bis w6_1 im Netzwerk-Graph und in den Bigram-Tabellen jeweils Wortgruppen, die unterscheidbar indiziert sind. Die Mitglieder einer Wortgruppe sind entsprechend durch Indizieren eindeutig zu kennzeichnen.

Fig. 3 veranschaulicht die Abfolge des Erkennungsprozesses für eine im Beispiel nach **Fig.** 2a, 2b als Satz zulässige Wortfolge w2w3w1. Ausgehend von einem Satzanfangsknoten Start sind als erstes Wort nur w1 oder w2 zulässig. Der Beginn eines Sprachsignals wird daher auf mögliche Übereinstimmung mit w1 und/oder w2 überprüft. Hierzu wird auf die in einem Aussprachelexikon L abgelegten sprachlichen Charakteristika dieser beiden Wörter zurückgegriffen. Gebräuchlicherweise enthalten die Lexikoneinträge zu jedem Wort mehrere Wortuntereinheiten mit Vorschriften über deren zulässige Aufeinanderfolge. Die Vorgehensweise bei der Worterkennung kann beispielsweise wie bei dem erwähnten Harpy-System durch Durchlaufen einer baumartigen Suchpfadstruktur erfolgen mit fortlaufender Bewertung der einzelnen untersuchten Pfade und Ausblenden von niedrig bewerteten Pfaden.

In **Fig.** 3 ist für die Suchstruktur vereinfacht eine lineare Kette mehrerer Wortuntereinheiten WU (Kreise) eingetragen.

Die Lexikoneinträge umfassen wie bereits erwähnt auch Wortendeknoten WE (Quadrate in **Fig.** 3), wobei für mehrfach an verschiedener Position im Graphen der **Fig.** 2a auftretende gleiche Wörter entsprechend deren Indizierung ein Lexikoneintrag mehrere Wortendeknoten aufweist, die jeweils einem der indizierten Exemplare desselben Worts durch den übereinstimmenden Index zuordenbar sind und die zulässigen Nachfolgewörter festlegen. Der Index eines Wortes wird beim Zugriff auf das Lexikon in der Weise berücksichtigt, daß mittels des Index die richtige Auswahl unter dem ggf. mehreren Wortendeknoten getroffen wird.

Bei dem in **Fig.** 3 skizzierten Beispiel ist angenommen, daß das Sprachsignal keine ausreichende phonetische Übereinstimmung mit dem Lexikoneintrag zum Wort w1 zeigt und dieser Teil des Suchpfads abgebrochen wird, noch bevor das Wortende von w1 erreicht ist. Hingegen zeige das Sprachsignal eine gute Übereinstimmung mit dem Lexikoneintrag zum Wort w2, so daß dieser Suchpfad weiterverfolgt wird. Da w2 im Sprachmodell nur an einer Position auftritt, existiert nur ein Wortendeknoten, von dem aus sich die Suche verzweigt auf die Überprüfung der Wörter w3 und w5 als zulässige Nachfolger, die erfindungsgemäß durch Indizieren als w3_2 und w5_1 eindeutig gemacht sind. Für w5 sei wieder mangelnde phonetische Übereinstimmung mit dem fortgesetzten Sprachsignal und Abbruch dieses Teils des Suchpfads angenommen, wogegen der Suchpfad über w3 bis zur Verzweigung auf die beiden Wortendeknoten mit Indizes _1 und _2 weiterverfolgt werde. Mittels des Index _2 aus dem indizierten Zugriff auf den Lexikoneintrag w3 wird der gleich indizierte Wortendeknoten für die Weiterführung des Suchpfads ausgewählt, woraus sich w1_2 als einziges zulässiges Nachfolgewort ergibt. Dessen Lexikoneintrag wird wieder mit dem fortgesetzten Sprachsignal verglichen. Bei ausreichender Übereinstimmung wird der Suchpfad über den mit _2 indizierten Wortendeknoten zum Satzende weiter geführt.

Im Realfall werden vorzugsweise mehrere Suchpfade vollständig bis zum Satzende verfolgt und danach einer weiteren Auswahl unterzogen, bei der beispielsweise durch Schwellwertsetzung und/oder Vergleich der globalen Satzwahrscheinlichkeiten oder anderer an sich bekannter Bewertungsgrößen einer der erkannten Sätze als bester Satz ausgewählt und weiter verarbeitet wird, z. B. als auszuführendes Kommando.

Patentansprüche

1. Verfahren zur Spracherkennung von aus mehreren Wörtern eines gegebenen Wortschatzes zusammengesetzten Sätzen, bei welchem eine begrenzte Menge zulässiger Sätze und ein Sprachmodell, in welches die Syntax der zulässigen Sätze integriert ist, vorgegeben wird, **dadurch gekennzeichnet,** daß für Wörter, die in der Menge der zulässigen Sätze mehrfach in verschiedenen syntaktischen Positionen auftreten, in dem Sprachmodell mehrfache und voneinander unterscheidbare Exemplare mit den für die jeweilige Position gültigen syntaktischen Einschränkungen vorgegeben werden, und daß durch fortlaufende Berücksichtigung der syntaktischen Einschränkungen des Sprachmodells während des laufenden Erkennungsprozesses nur die Übereinstimmung eines aktuellen Sprachsignals mit zulässigen Wortfolgen überprüft wird.
2. Verfahren nach Anspruch 1, gekennzeichnet durch einen HMM-Erkennungsprozeß.
3. Verfahren nach Anspruch 1 oder 2, dadurch gekennzeichnet, daß als Sprachmodell ein Bigram-Modell vorgegeben wird.

4. Verfahren nach einem der Ansprüche 1 bis 3, dadurch gekennzeichnet, daß den im Sprachmodell mehrfach vorgegebenen Exemplaren eines Wortes derselbe Eintrag in einem Aussprachelexikon zugewiesen wird, der durch eine Auswahl von Wortendeknoten eindeutig einem der mehreren Exemplare zugeordnet wird.

5. Verfahren nach einem der Ansprüche 1 bis 4, dadurch gekennzeichnet, daß den bei dem Erkennungsprozeß überprüften zulässigen Wortfolgen Wahrscheinlichkeitswerte zugewiesen und diese einem Schwellwertvergleich unterzogen werden.

6. Verfahren nach Anspruch 5, dadurch gekennzeichnet, daß als Wahrscheinlichkeitswert die globale Wortfolgenwahrscheinlichkeit, normiert auf die aktuelle Wortfolgenlänge ermittelt wird.

Hierzu 3 Seite(n) Zeichnungen

– Leerseite –

Nummer: DE 195 01 599 C1
Int. Cl.[6]: G 10 L 5/06
Veröffentlichungstag: 2. Mai 1996

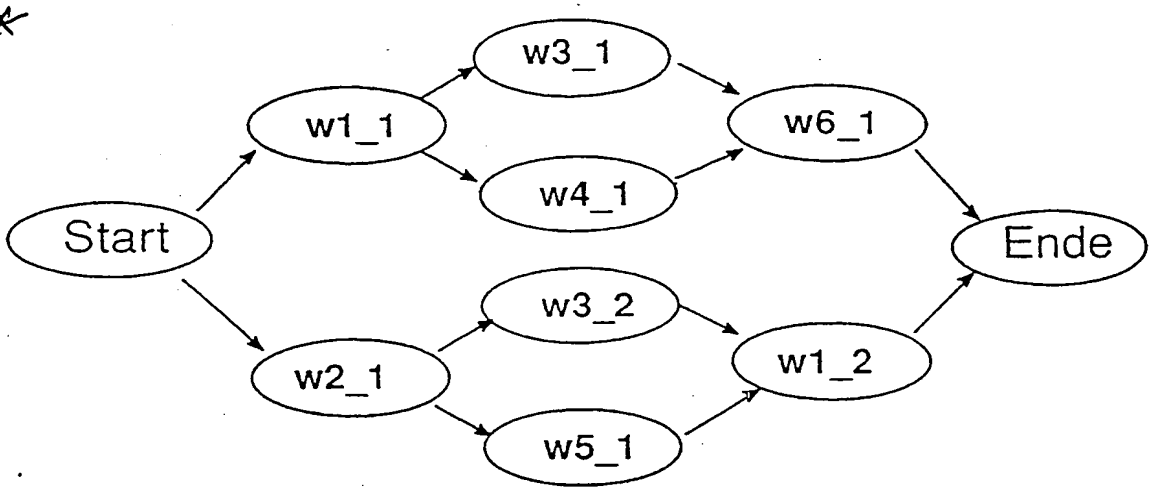

FIG.2a

Wort	zulässige Nachfolger
Start	w1_1, w2_1
w1_1	w3_1, w4_1
w1_2	Ende
w2_1	w3_2, w5_1
w3_1	w6_1
w3_2	w1_2
w4_1	w6_1
w5_1	w1_2
w6_1	Ende

FIG.2b

Nummer: DE 195 01 599 C1
Int. Cl.[6]: G 10 L 5/06
Veröffentlichungstag: 2. Mai 1996

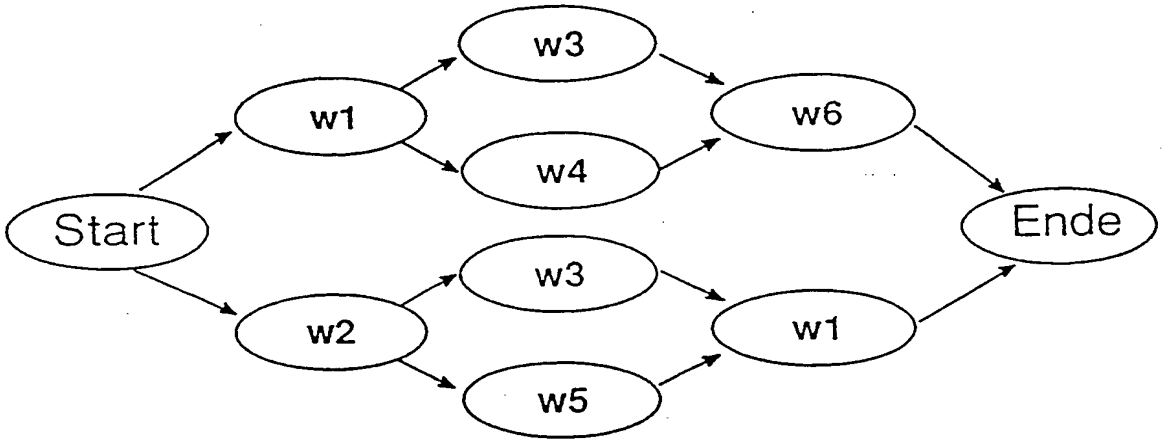

FIG.1a

Wort	zulässige Nachfolger
Start	w1, w2
w1	w3, w4, Ende
w2	w3, w5
w3	w1, w6
w4	w6
w5	w1
w6	Ende

FIG.1b

Nummer: DE 195 01 599 C1
Int. Cl.[6]: G 10 L 5/06
Veröffentlichungstag: 2. Mai 1996

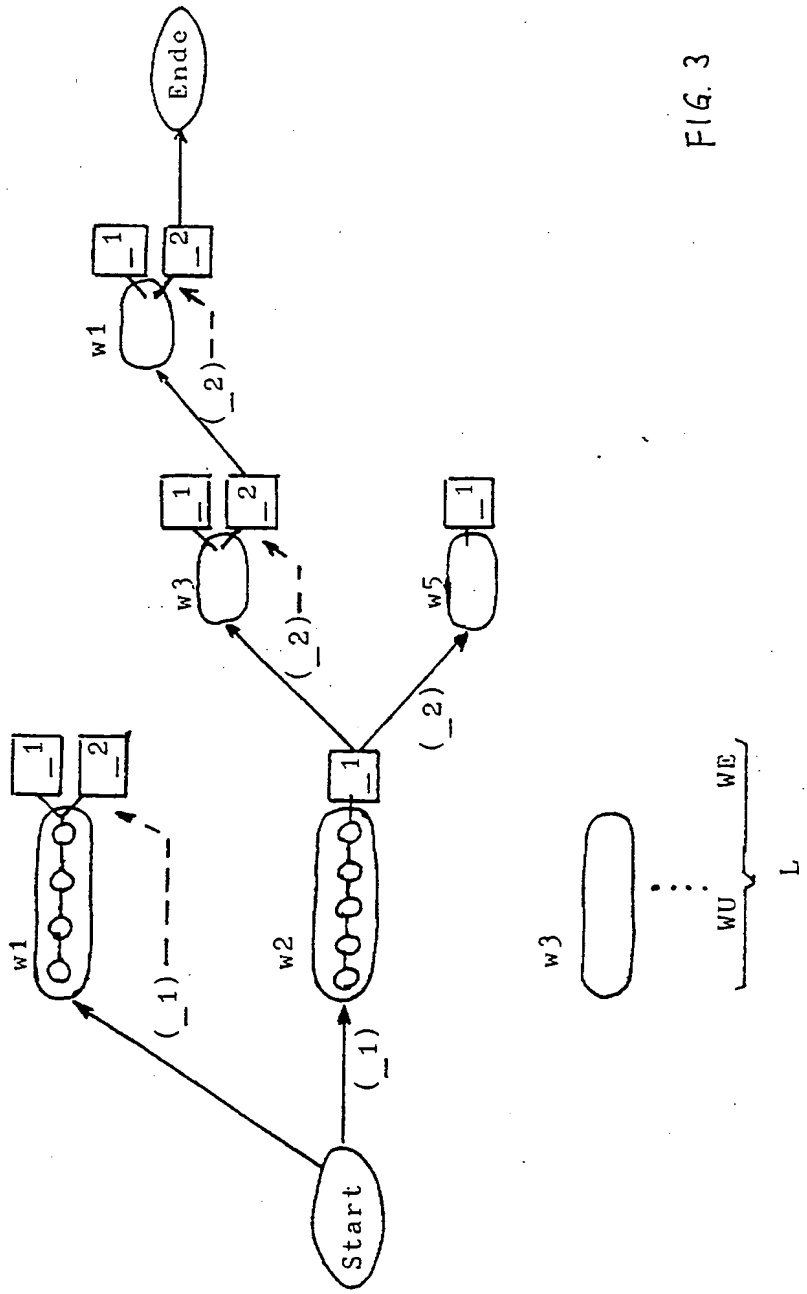

FIG. 3